Low latency and tight resources viseme recognition from speech using an artificial neural network

Nathan Souviraà-Labastie , Frédéric Bimbot

RESEARCH REPORT

N° 8338

26/07/2013

Project-Team PANAMA

ISSN 0249-6399

Low latency and tight resources viseme recognition from speech using an artificial neural network

Nathan Souviraà-Labastie[1] , Frédéric Bimbot [2],
Project-Teams PANAMA

Research Report N° 8338 — 26/07/2013 —15 pages.

Abstract: We present a speech driven real-time viseme recognition system based on Artificial Neural Network (ANN). A Multi-Layer Perceptron (MLP) is used to provide a light and responsive framework, adapted to the final application (*i.e.*, the animation of the lips of an avatar on multi-task platforms with embedded resources and latency constraints). Several improvements of this system are studied such as data selection, network size, training set size, or choice of the best acoustic unit to recognize. All variants are compared to a baseline system, and the combined improvements achieve a recognition rate of 64.3% for a set of 18 visemes and 70.8% for 9 visemes. We then propose a tradeoff system between the recognition performance, the resource requirements and the latency constraints. A scalable method is also described.

Key-words: Speech Processing, Lip Animation, Visemes, Artificial Neural Network, Computational Cost.

[1] IRISA/Université de Rennes 1 – nathan.souviraa-labastie@irisa.fr

[2] IRISA/CNRS UMR 6074 – frederic.bimbot@irisa.fr

RESEARCH CENTRE
BRETAGNE ATALANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex France

Reconnaissance de visèmes à partir du signal de parole basé sur un réseau de neurones répondant à des contraintes de latence et de coût de calcul

Résumé : Ce rapport présente un système de reconnaissance de visèmes à partir du signal de parole utilisant un réseau de neurones artificiels et capable de fonctionner en temps réel. Un Multi-Layer Perceptron (MLP) permet d' obtenir une méthode rapide et légère adaptée à l' application finale (i.e., l' animation des lèvres d' un avatar par une plateforme multitâche de type set-top-box avec des contraintes de ressources et de latence). Plusieurs améliorations de ce système sont également présentées telles que la sélection des données d' apprentissage, la taille du réseau, la taille de la base d' apprentissage ou encore le choix de l' unité acoustique à reconnaître. Toutes ces variantes sont comparées au système de base. La combinaison de toutes ces améliorations permet d' atteindre un taux de reconnaissance de 64.3% pour un jeu de 18 visèmes et 70.8% pour 9 visèmes. Nous proposons ensuite un système faisant le compromis entre performance, besoin en ressources et latence. Une variante adaptable (scalable) est aussi décrite.

Mots clés : Traitement de la parole, Animation labiale, Visèmes, Réseau de neurones, Ressources de calcul

1. Introduction

The visual component of speech provides valuable information that undoubtedly increases its intelligibility. This property is exploited in the design of synthetic animated faces, with which it is possible to obtain a much more natural aspect and interaction in many man-machine communication situations. For avatar animation, the main task consists of synchronizing the lip movements of the virtual face with the speech signal of a human subject driving his/her avatar.

As a consequence of the current high-speed network development, avatar animation in general and animated faces in particular is a growing research field, with rapidly expanding potential applications. Video games, eLearning, instant messaging or 3D animation productions (movies, commercials, *etc.*) need also to generate lip animations. As part of the ReV-TV[3] project, the avatar represents a television-viewer in a new kind of TV programs. In this case, the implementation on a set-top-box leads to low resources and latency constraints. Other embedded multi-tasks platform, such as mobile phones are targeted.

In practice, the lip movements are monitored with a discrete set of lip positions: the visemes. The extraction of visemes may imply the detection of phoneme sequences (*e.g.*, [1]) which are then mapped to visemes (*i.e.*, a lip configuration of the animated face). Alternatively, the recognition system can also be trained to directly recognize visemes (*e.g.*, [2], [3]). Hidden Markov Models (HMM) [3], ANN [2], [4], Gaussian Mixture Models (GMM) or Vector Quantization can be used to map acoustic vectors to the corresponding sequence of acoustic units.

In this paper, we introduce a real-time viseme recognition system. We have therefore focused our efforts on an ANN: the MLP. We study its ability to match audio features with lip positions, without resorting to conventional speech recognition techniques (in particular, HMM), which use high-level linguistic information, and would introduce too much delay in the viseme estimation process.

In section 2, the architecture of the system is described before more details about the training and testing setup are given in the third section. Section 4 presents experiments that show the improvements in terms of recognition rates provided by several approaches. Then, the advantages and disadvantages of the resources and latency of this approaches are discussed, and a tradeoff system is presented in section 5. Section 6 compare this tradeoff system with state of the art and present the perspectives.

2. Architecture of the system

The real-time speech driven system used in this work is illustrated in Figure 1. It is based on a simplified decomposition [5] of Automatic Speech Recognition (ASR) into sequential modules. Given the constraints presents within the application, the recognition process must be causal and should therefore not use any lexical (such as phonetic trees) or syntactical information (such as language models).

[3] http://www.rev-tv.eu/

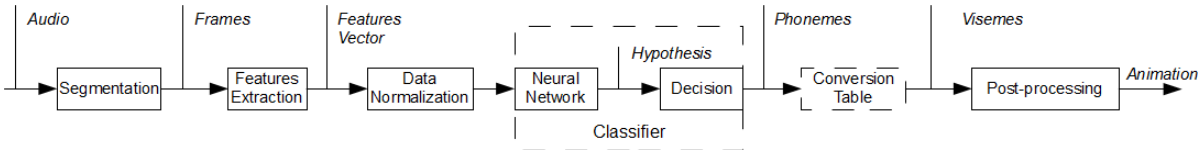


Figure 1: Illustration of the speech driven lip animation system.

The system is composed of 7 steps. The audio signal is first segmented in 20 ms frames with 10 ms overlap between frames. This overlapped segmentation provides a good representation of the signal and is commonly used in state of the art. The audio features extracted in the second module are the MFCC coefficients (13 static coefficients for the baseline system). In the third module, the audio features are normalized. The normalization step balances the range of the different features at the input of the classifier. The chosen neural network (4th module) is a MLP with one hidden layer. The decision module (5th in the sequence) selects the output of the network according to the maximal value. If the classifier recognizes phonemes, a conversion module from phonemes to visemes (6th module) would be added to the system (see Table 1). The post-processing module (7th module) aims at improving (in real-time) the final animation. It needs at least a 30 ms look-ahead length to provide a more coherent output sequence of visemes.

3. Experimental setup

3.1. Speech data

The speech database used to train the neural network is a subset (5 hours) of the ESTER database [6]. ESTER is a multi-speaker database of radio program which has been phonetically annotated by an automatic system. A set of 36 phonemes is used to describe the French language, plus silence and inspiration labels (*cf.* Table 1). They have been aggregated in two systems of viseme classes. One is composed of 18 classes, based on Benoît et *al.* [7], and has been reduced into 9 classes of visemes according to Govokhina works [8].

Each training example corresponds to an audio frame, represented as its feature vector and its phonemic or visemic transcription. The training set is roughly composed of 200 000 examples stemming from the 5 hours training data set and the test set contains 100 000 other frames (from 30 other minutes of ESTER). The difference between the ratios is caused by the data selection used for generating the training set (*cf.* section 4.2.).

3.2. Frame-by-frame evaluation

The test set is strictly different from the training sets. The various configurations of the system presented in the experiments section have been evaluated on the same subset of 30 minutes taken from another part of the database (*cf.* § 3.1.). Recognition rates are thus fully comparable. Depending on what the network is trained to recognize, the Phoneme or the Viseme Recognition Rate (PRR or VRR) are used.

VIS	Phonemes (18 visemes)	Phonemes (9 visemes)	VIS
V0	silence	silence	V0
V1	inspiration	inspiration	V1
V2	[p][b][m]	[p][b][m]	V2
V3	[f][v]	[f][v]	V3
V4	[ʃ][ʒ]	[ʃ][ʒ]	V4
V5	[e][ɛ][œ̃][ɛ̃]	[e][ɛ][œ̃][ɛ̃][a][ɑ][i][j]	V5
V6	[a][ɑ]		
V7	[i]		
V8	[j]		
V9	[ɔ]	[ɔ][ɑ̃]	V6
V10	[ɑ̃]		
V11	[t][d][n][ɲ]	[t][d][n][ɲ][s][z][k][g][l][ʁ]	V7
V12	[s][z]		
V13	[k][g]		
V14	[l]		
V15	[ʁ]		
V16	[y][u][ə][ø][œ][o][ɔ̃][ɥ]	[y][u][ə][ø][œ][o][ɔ̃][ɥ] [w]	V8
V17	[w]		

Table 1: Conversion table (phonemes into visemes). The phonetic symbols are taken from the International Phonetic Alphabet (IPA).

3.3. Neural network topology

We have focused this work on Regular MLPs with a sigmoidal transfer function because of their low-complexity and fast-responding abilities. They have been considered, as they constitute a reasonable reference for classification tasks, but other topologies can be envisaged such as Time-Delay Neural Network (TDNN) or Recurrent Neural Network (RNN). This choice is discussed more deeply in section 5.2.

4. Experiments

This section presents several approaches used to optimize the configuration of the previous system. The objective is to increase the PRR and the VRR of this baseline system by using state of the art techniques in addition to those proposed in this paper. Table 2 presents results obtained from this successive improvements with regards to their respective impact on latency and model size.

In order to assess the relevance of a non-linear classifier, a 1 hidden layer MLP is first compared to a linear model (implemented as a MLP with no hidden layer). Table 2 shows unquestionably that the MLP has a significantly higher recognition rate than a linear classifier.

All other reported results are obtained with MLPs with one hidden layer. Table 2 also gives two other characteristics: the number of free parameters of each MLP (#param.) and the look-ahead length. The first one corresponds to the number of biases and weight of the network and sometimes input normalization parameters. The second one indicates how long the system delays the animation compared with the input speech (latency).

			38 Phonemes		18 Visemes		9 Visemes	
System	Added improvements	Look-ahead length	PRR	#param.	VRR	#param.	VRR	#param.
0	Linear classifier	30 ms	33.0%	538	38.7%	538	53.4%	538
1	MLP - 35 nodes	30 ms	39.9%	1858	46.8%	1858	58.3%	1858
2	Data selection by temporal center	30 ms	44.0%	1858	50.7%	1858	62.5%	1858
3	Optimal size of the network and the train set	30 ms	47.1%	4458	53.1%	4458	64.5%	4458
4	Data normalization	30 ms	48.9%	4484	54.8%	4484	66.1%	4484
Tradeoff	Direct Visemes Recognition (DVR)	30 ms	--	--	55.9%	2764	66.9%	1990
6	Use of the first and second features derivative	70 ms	--	--	64.3%	5026	70.8%	4252

Table 2: Summary of the progressive addition of the improvement (systems 0 to 5 are based on the VVP approach).

4.1. Balancing the training set

Two ways of balancing the number of training examples per classes have been considered. The first type of repartition enforces an identical number of training examples for each class of phonemes or visemes. The second gives a proportional weight to classes depending on the phonetic repartition of the language (determined, for instance, on the training data).

The second approach turns out to provide a better performance. The remaining experiments presented here use this language repartition for balancing the training set.

4.2. Improving data selection for training

We investigated 4 ways to summarize the acoustic content of a phonetic segment within a single feature vector.

The temporal center and the centroid selection can be used to select a representative frame of each occurrence of a phoneme: (i) the temporal center selection keeps the central frame of the phoneme; (ii) the centroid selection extracts the frame which is closest to all other frames within the phonetic segment. The mean (iii) and the median (iv) approaches consist in computing, for each phoneme occurrence, a feature vector as the mean or the median of each feature considered separately. Even though it is quite an unusual approach, the median approach has been considered because of its possible robustness to noisy environments.

As can be seen in Table 3, the data selection process has a visible impact on the recognition performances. In particular, the temporal center selection performs among the best and is therefore retained as a simple and natural way of selecting robust training data. It is worth noting that these data selections clearly improve the recognition rates without increase the training data size, the number of parameters (fixed hidden layer size), or the look-ahead length (*cf.* Table 2). It constitutes one of the main contribution of this paper.

Data selection	Phoneme Recognition Rate (PRR)
All the frames	39.9%
Temporal center (i)	44.0%
Centroid (ii)	42.7%
Mean (iii)	42.3%
Median (iv)	44.2%

Table 3: Results of each type of training data selection.

4.3. Optimal network size

As a general trend, the more training examples are available to a classifier, the more accurately the decision borders can be estimated. In our experiments, the number of training examples is limited to 200 000 (*i.e.*, one hour of speech) because of the computer memory.
A rule of thumb [9] says that 10 times more examples than free parameters are needed to ensure a good learning. However, in this study, 85 hidden nodes in the hidden layer , *i.e.*, about 5000 free parameters, has been found to be a satisfactory size above which classification performance plateaued. This indicates that, in our experiments, the optimal ratio between training examples and degrees of freedom is rather in the order of 40. The optimization of the network size increases the performance by 3%.

4.4. Normalization

In order to give a comparable weight to feature vectors components, their values are normalized : the mean is set to zero and variance to one. Training and test data undergo the transformation estimated on the training set. This form of normalization is widely used in speech recognition and lead to a performance increase of 2%.

4.5. Acoustic unit selection

One architectural choice is to decide at which step the phoneme-to-viseme conversion module should take place. Two approaches are compared: the "Viseme-Via-Phoneme" (VVP) approach where the visemes are obtained after a phoneme recognition step via a conversion table, and the "Direct Viseme Recognition" (DVR) where the training samples are previously converted into visems and the MLP directly outputs a viseme hypothesis.
The last line of Table 2 corresponds to the DVR configuration, whereas all others correspond to a VVP approach. The DVR approach brings 4% of improvement for the 18 viseme classes, but only 0.7% for the 9 viseme classes.

4.6. Use of dynamic features

Schwarz *et al.* [11] report a performance benefit on using the first two derivatives of the MFCC coefficients. Two configurations routinely used in speech recognition are compared: one with 13 static MFCC coefficients and one also using the first two derivative features (39 coefficients).
As expected, the second configuration has shown an advantage in terms of performance (*cf.* Table 2). On the other hand, the look-ahead length and the number of parameters increase.

5. Latency and low resources constraints

In the previous section, gradual improvements in the configuration provide an overall performance gain typically in the range of 14% absolute performance, and have placed us in an average range of performance comparing to state of the art (*cf.* Table 4). It has to be noted that the recognition performance are not fully comparable because of the different using tasks and databases. However, some improvements of the process cause an increase of required resources and look-ahead length. In this part, we first discuss about these constraints stemming from the application, and then present in detail the tradeoff system.

5.1. Latency

The baseline system has a 30 ms look-ahead length due to the post-processing, and 40 ms more are added to reap the benefits of MFCC derivatives. In our case, the audio and the animation is synchronized according to the total latency.

We also need to preserve the flow of conversation. ITU [11] recommends a maximum 150 ms delay between the speaker and the listener. But, the telecommunication network latency is unpredictable. So, a system with minimum latency has to be inserted. Adding 50 ms of delay seems to be an upper limit without definitely affecting the flow of conversation.

According to the latency caused by the communication network, a tradeoff has to be found between the quality of the animation (represented by the recognition rate) and the latency inserted into the flow of conversation. As an adaptive solution, a scalable system could be implemented to deal with the communication network latency instability. For this, the tradeoff system and the system 6 (*cf.* Table 4) could be used alternatively, depending on the communication network latency.

Systems			Performances			Costs		
References / Real-Time / Classifier			Data Type / Recognition Rates / #classes			Look-ahead length / #param / #mul. per frame		
System 4	Y	MLP	Radio	48.9%	38 phonemes	30 ms	4k	7k
Salvi [4]	Y	RNN	Telephone	54.2%	38 phonemes	60 ms	541k	(>540k)
Salvi [4]	Y	RNN-Viterbi	Telephone	55.3%	38 phonemes	(160 ms)	541k	(>540k)
Tradeoff	Y	MLP	Radio	55.9%	18 visemes	30 ms	3k	3k
Bozkurt [3]	N	GMM/HMM	TIMIT	60.1%	16 visemes	>30 ms	(11k)	(>11k)
System 6	Y	MLP	Radio	64.3%	18 visemes	70 ms	5k	6k
Bozkurt [3]	N	GMM/HMM	TIMIT	73.0%	16 visemes	>30 ms	(114k)	(>113k)
Massaro [2]	Y	TDNN	Telephone	46.0%	9 visemes	(140 ms)	(>108k)	(>54k)
Tradeoff	Y	MLP	Radio	66.9%	9 visemes	30 ms	2k	3k
System 6	Y	MLP	Radio	70.8%	9 visemes	70 ms	4k	5k
Luo [15]	Y	RNN	TIMIT	84.7%	9 visemes	40 ms	(5k)	(6k)

Table 4: Comparison to state of the art. Our own system's references are given in boldface. Systems are grouped by comparable Recognition Rates. The values in brackets have been inferred from the descriptions in [2], [3], [4], and [15].

5.2. Resources

The computational cost is expressed in terms of multiplications per frame (mpf), *i.e.*, per 10 ms. The sums are ignored insofar as the calculations can be done on a DSP [12].

As shown in [13], the MFCC extraction can be reduced to 804 mpf. The MLP computation needs transfer functions, multiplications and sums. Using a look-up table (*e.g.*, [14]) to approximate the sigmoidal transfer function, no multiplication are added. Regarding the memory storage, four bytes for each parameter in memory can be considered.

As an example, the entire final system that recognized 9 classes of visemes (system 6 in Table 2 and 4) has a complexity of 4923 mpf (804 for the MFCC extraction and 4119 for the MLP) and used 17 kB to store the parameters. Considering architecture as in [12], [13], and [14], the latency stemming from this system (far less than 1ms) is much lower than the look-ahead length and can thus be ignored.

As previously explained, this method is aimed at embedded systems, on which no dedicated architecture is available. The task is almost continuous and will be hosted on multi-task platforms, it therefore needs to run as an idle task and to require little computational resources in order to facilitate its scheduling. This lead to the choice of MLPs instead of TDNNs [2] and RNNs [4], [15] which need to store former state of

neurons, or GMM/HMM [3] which imply multiple levels of analysis. In addition, the number of parameters used by the neural network and the number of multiplication per frame are also reduced.

5.3. The tradeoff system

The dynamic features are excluded from the tradeoff system owing to their resource requirement and the added latency. All the other improvements does not add as much cost as the dynamic features, and are kept to contribute to the tradeoff system. The recognition performance is then slightly lower (3.9% of absolute VRR for 9 visemes classification), but the key constraints are divided by two (*cf.* Table 4).
For 9 classes classification, this tradeoff system is composed of 1990 parameters (8 kB), and requires 2687 mpf (804 for the MFCC extraction and 1883 for the MLP).

6. Conclusion and perspectives

This article has presented several experiments for improving viseme recognition of an on-line lip animation system. Experiments that have provided the best improvements are the data selection for training, the use of derivative features, and the "Direct Viseme Recognition" approach. Comparisons with similar tasks (*e.g.*, [2], [15]) would seem to place our system in an average range of performance in terms of recognition rate but these comparisons are difficult to interpret because of the differences between experimental conditions (*e.g.*, quality of the learned speech database, use of contextual information), and the targeted applications.

However, it is important for the targeted applications to address the question of resources and latency. Several system configurations have been given with their recognition rates and resource requirements. Then, a very light tradeoff system that fitted our specific needs has been detailed. It has a 30 ms look-ahead length from speech to animation, with less than 2000 parameters and very light impact on architecture. Thus, competitive systems have been proposed in terms of latency comparing to [16], resource requirements comparing to [17], and both for [2], [3], [4], and [15], while first tests give satisfaction on visual result. We also proposed to use them in a scalable fashion to address the problem of the communication network latency instability.

Future work will aim at improving the current approach and comparing it with a bi-parametric (horizontal and vertical) representation of the lips, possibly recognized by an ANN. A specific database [18] has been collected for this purpose.

7. References

[1] T. Kim *et al.*, "Achieving real-time lip synch via SVM-based phoneme classification and lip shape refinement", *Proc. ICMI'02*, pp. 299–304, 2002.

[2] D. W. Massaro *et al.*, "Picture my voice: Audio to visual speech synthesis using artificial neural networks", *Proc. AVSP'99*, 1999.

[3] E. Bozkurt *et al.,* "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation", *3DTV Conference*, pages 1–4, 2007.

[4] G. Salvi, "Truncation error and dynamics in very low latency phonetic recognition", *ISCA workshop on Non-linear Speech Process.*, 2003.

[5] R. Boite et al,. "Traitement de la Parole", Presses Polytechniques Universitaires Romandes, Lausanne, 2000.

[6] G. Gravier *et al.,* "The ESTER evaluation campaign of Rich Transcription of French Broadcast News", *Proc. Language Evaluation and Resources Conference*, 2004.

[7] C. Benoit *et al.*, "A set of French visemes for visual speech synthesis", *Les cahiers de l'ICP*, Vol. 3, pp. 113-129, 1994.

[8] O. Govokhina, "Modèles de trajectoires pour l'animation de visages parlants", Thèse de l'INP de Grenoble, 2008.

[9] K. Messer and J. Kittler, "Choosing an optimal neural network size to aid a search through a large image database", *Proc. of the Ninth British Machine Vision Conference*, 1998.

[10] P. Schwarz *et al.*, "Hierarchical Structures of Neural Networks for Phoneme Recognition", *Proc. ICASSP*, pp. 325–328, 2006.

[11] Rec. ITU-T G.114, "One-way transmission time", 2003.

[12] R. Battiti *et al.,* "Special-purpose parallel architectures for high-performance machine learning", *High Performance Computing and Networking*, Milano, Italy, 1995.

[13] W. Han *et al.,* "An efficient MFCC extraction method in speech recognition", *Proc. IEEE Int. Symp. on Circuits and Syst. (ISCAS)*, p. 4, 2006.

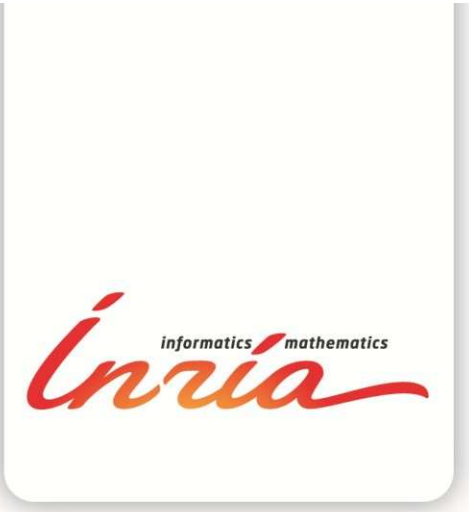
[14] P. K. Meher, "An optimized lookup-table for the evaluation of sigmoid function for artificial neural networks", *VLSI System on Chip Conference (VLSI-SoC)*, 2010.

[15] S.-H. Luo and R. W. King, "A novel approach for classifying continuous speech into visible mouth-shape related classes", *Proc. ICASSP*, Vol. 1, pp. 465-468, 1994.

[16] Z. Wen *et al.*, "Real time speech driven facial animation using formant analysis", *Proc. of ICME*, pp. 817-820, 2001.

[17] J. Park and H. Ko, "Achieving a reliable compact acoustic model for embedded speech recognition system with high confusion frequency model handling", *Speech Communication*, vol. 48, no. 6, pp. 737–745, 2006.

[18] Y. Benezeth *et al.*, "BL-Database: A French audiovisual database for speech driven lip animation systems", Inria Research Rep., 2011.



RESEARCH CENTRE
BRETAGNE ATALANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex France

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr
ISSN 0249-6399